Metric Measurement Invariance of Latent Variables: Foundations, Testing, and Correct Interpretation

Stefan Klößner
Statistics and Econometrics
Saarland University

Eric Klopp
Department of Education
Saarland University

Preliminary Version: April 27, 2017

Abstract

In multi-group and longitudinal studies, it is important to test for metric measurement invariance. Recently, several authors have pointed out that currently used test procedures for measurement invariance (MI) do not fully test for MI and that additional assumptions about the invariance of the referent indicator are needed in order to conclude that actual data satisfy MI.

Introducing the new concept of proportional factor loadings (PFL), we show that tests for MI actually only test for PFL, because PFL is empirically indistinguishable from metric MI. More precisely, if the loadings are only proportional over groups or time, the implied distribution of the observed variables is identical to one that results from invariant factor loadings. Thus, it is impossible to differentiate between MI and PFL based on empirical data only. Furthermore, PFL affects tests about the equality of latent variables' variances, leading to wrong conclusions when the data only satisfy PFL, but not MI.

We also discuss how the empirical indistinguishability between PFL and MI affects partial MI. We find that it is typically impossible to differentiate invariant from non-invariant indicators. Empirically, one can only detect which indicators form subsets whose loadings are proportional. These findings explain why procedures for detecting invariant indicators perform poorly under certain conditions, a disturbing fact that several recent studies have found in Monte Carlo studies.

Finally, we also discuss the referent indicator problem and how different scaling methods potentially affect the results of the above mentioned tests.

1 Introduction

In multi-group and longitudinal studies, it is important to test for metric measurement invariance. Correspondingly, this topic has spurred a huge number of publications since the beginnings with the seminal papers of Byrne et al. (1989) and Meredith (1993).[1] The importance of measurement invariance stems from the fact that conclusions about a latent variable's behaviour in different groups based on observed variables' variances and covariances may be severely wrong when measurement invariance is lacking. Therefore, there is largely consensus in the literature that before drawing such conclusions, a thorough investigation of measurement invariance is mandatory: only if the latent variable measures, at least to a large degree, the same underlying construct in all groups, comparisons between groups are sensible.

Recently, several authors have pointed out that currently used test procedures for measurement invariance (MI) do not fully test for MI and that additional assumptions about the invariance of the referent indicator are needed in order to conclude that actual data satisfy MI. Introducing the new concept of proportional factor loadings (PFL), we show that tests for MI actually only test for PFL, because PFL is empirically indistinguishable from metric MI. More precisely, if the loadings are only proportional over groups or time, the implied distribution of the observed variables is identical to one that results from invariant factor loadings. Thus, it is impossible to differentiate between MI and PFL based on empirical data only.

Furthermore, PFL affects tests about the equality of latent variables' variances, potentially leading to wrong conclusions when the data only satisfy PFL, but not MI. Using Monte Carlo studies, we show that this problem is so severe that tests may have no power at all to detect diverging variances or may with a probability of 100% wrongly indicate that latent variances were different, although the data actually come from a process with identical latent variances.

We also discuss how the empirical indistinguishability between PFL and MI affects partial MI. We find that it is typically impossible to differentiate invariant from non-invariant indicators. Instead, empirically one can only detect which indicators form subsets whose loadings are proportional. As we show, those subsets can be detected either by estimating the underlying model using different scaling methods, or alternatively by calculating properly chosen ratios of factor loadings and investigating these with respect to invariance across groups.

Our paper is structured as follows: in Section 2, we introduce the notion of proportional factor loadings (PFL) and discuss how tests on metric measurement invariance fail to detect PFL. Building on this finding, section 3 theoretically shows that it is impossible to tell apart

[1] See Putnick and Bornstein (2016) for a very recent overview over the literature and the state of the art regarding measurement invariance.

PFL and metric MI based on data only, and that testing for metric MI is actually tantamount to testing for PFL. Section 4 is devoted to the consequences of the empirical indistinguishability of PFL and MI, in particular it is concerned with the question of how to properly interpret a potential non-rejection of metric MI. In section 5, we discuss how PFL may distort the testing for equality of latent variances across groups, while section 6 is devoted to partial MI and the distinction betwenn 'proportional' and 'invariant sets' of indicators. Eventually, section 7 concludes.

2 Metric Measurement Invariance & Proportional Factor Loadings

Throughout this paper, we will consider the following well-known model for testing metric measurement invariance in the context of confirmatory factor analysis (CFA):[2] some latent variables ξ, also called common factors in CFA, are indirectly measured through observed variables X, being linked by the following equations that hold in every group g:

$$X_g = \tau_g + \Lambda_g \xi_g + \delta_g, \tag{1}$$

where τ_g, Λ_g, and δ_g denote intercepts, factor loadings, and measurement errors in group g, respectively. Denoting the latent and observed variables' mean by the symbols α and μ, equation (1) implies the following relation:

$$\mu_g = \tau_g + \Lambda_g \alpha_g, \tag{2}$$

while the variances and covariances are linked via

$$\Sigma_g = \Lambda_g \Phi_g \Lambda_g' + \Theta_g, \tag{3}$$

where α_g and μ_g denote the latent and observed variables' covariances, while Θ_g denotes the error term's covariance matrix in group g. Measurement invariance (MI) is concerned with the question whether certatin of these quantities are invariant over groups: for instance, one speaks of metric MI when Λ_g is identical for all groups $g = 1, \ldots, G$: $\Lambda_1 = \ldots = \Lambda_G$. When, additionally, the intercepts are invariant, too ($\tau_1 = \ldots = \tau_G$), one speaks of scalar invariance. In this paper, however, we will restrict ourselves to investigating the problems associated with testing for metric MI, leaving for future research a thorough investigation of

[2] We adopt the notation of Yoon and Millsap (2007).

the problems appearing when testing for scalar MI.

In the literature, one particular case of metric non-invariance has attracted special attention: the scenario where the non-invariant factor loadings in a specific group (typically, the second group in a two-group setting) are uniformly lower (or higher) than in the other groups. This case is often termed 'uniform pattern of non-invariance' and it has been investigated by, i.a., Meade and Lautenschlager (2004), Meade and Bauer (2007), Chen (2007), Yoon and Millsap (2007), Chen (2008), French and Finch (2008), and Whittaker and Khojasteh (2013). With respect to this scenario, the literature has found two main results: first, the power of detecting non-invariance is considerably lower as compared to a scenario where the non-invariant loadings are both larger and smaller than those in other groups. While this result may not be surprising, the second finding is intriguing and rather counter-intuitive: under uniform non-invariance of loadings, the power to detect non-invariant loadings may decrease when the number of non-invariant indicators grows larger.

A particular special case of uniform non-invariance is the one discussed by Yoon and Millsap (2007), where non-invariant factor loadings are given by proportional rescalings: for instance, in one of their simulation settings, non-invariant values in the second group were all constructed as 4/7 of the corresponding loadings in the first group. In line with the above cited literature, Yoon and Millsap (2007) find that the power to detect non-invariance deteriorates drastically when more than half of the factor loadings are non-invariant. Carrying this idea to the extremes, we will now discuss the case where *all* factor loadings are proportional between the groups, but not invariant. We term this case 'proportional factor loadings' (PFL), formally it is defined by postulating that the loadings are identical up to some constant factors:

$$\text{(PFL): there exist constants } c_1, \ldots, c_G \text{ such that } c_1\Lambda_1 = \ldots = c_G\Lambda_G. \tag{4}$$

Obviously, metric MI implies PFL, while PFL may be fulfilled although metric MI is not. Thus, ideally, we would like tests for metric MI to be able to reliably detect PFL as a violation of metric MI. Speaking in statistical terms, we would like tests of metric MI to have large power for telling apart metric MI from PFL. In order to investigate this issue, we built on the high-frequency, large loading differences setup of Yoon and Millsap (2007) and conducted a small Monte Carlo study: we simulated data for two groups of size 500 each, employing a most simple one-factor model with six indicators, with loadings of $(.7, .9, .5, .6, .8, .3)$ and $4/7 \cdot (.7, .9, .5, .6, .8, .3)$ in the first and second group, respectively. Factor variances were 1 and 1.3, respectively, while error term variances were identical in both groups, given by $(.7, 1.2, .4, .5, .9, .2)$. Replication size was 1,000, with different seeds for initializing the random

generator of R[3] and the simulation conducted using package lavaan[4]. Using the well-known likelihood ratio test which compares the χ^2-values of the unrestricted model without equality conditions to that of the restricted model postulating $\Lambda_1 = \ldots = \Lambda_G$, we found that at the 5% significance level,[5] metric MI was rejected in only 57 out of the 1,000 cases.[6] This empirical detection rate of 5.7% is extremely low and very close to the test's nominal significance level of 5%, thus it seems that the test has no power at all to tell apart PFL from metric MI.[7] In the sequel, we will therefore investigate theoretically why it is impossible to detect violation of metric measurement invariance when the data generating process is characterized by proportional factor loadings.

3 Testing for Metric Measurement Invariance is Tantamount to Testing for Proportional Factor Loadings

Before attacking the general case, we will first explain why there is no power do tell apart PFL from metric MI in the example considered above. To this end, recall that the loadings were given by $\Lambda_1 = (.7, .9, .5, .6, .8, .3)'$ and

$$\Lambda_2 = 4/7 \cdot (.7, .9, .5, .6, .8, .3)' = (0.4, 0.5143, 0.3429, 0.2857, 0.4571, 0.1714)'$$

in the first and second group with factor variances of $\Phi_1 = 1$ and $\Phi_2 = 1.3$, respectively. This implies that the term $\Lambda_2\Phi_2\Lambda_2'$ appearing in Equation (3) is given by

$$\Lambda_2\Phi_2\Lambda_2' = \begin{pmatrix} 0.2080 & 0.267 & 0.1783 & 0.1486 & 0.238 & 0.0891 \\ 0.2674 & 0.344 & 0.2292 & 0.1910 & 0.306 & 0.1146 \\ 0.1783 & 0.229 & 0.1528 & 0.1273 & 0.204 & 0.0764 \\ 0.1486 & 0.191 & 0.1273 & 0.1061 & 0.170 & 0.0637 \\ 0.2377 & 0.306 & 0.2038 & 0.1698 & 0.272 & 0.1019 \\ 0.0891 & 0.115 & 0.0764 & 0.0637 & 0.102 & 0.0382 \end{pmatrix}.$$

[3] See R Core Team (2016).

[4] See Rosseel (2012).

[5] We conducted the corresponding test using in turn each of the six indicators as referent variable, and additionally also using the effects coding method of Little et al. (2006). The choice of the particular scaling method has no influence on the test results, see also below.

[6] The literature provides alternatives to the χ^2-differences test for testing metric MI, see e.g., Cheung and Rensvold (2002), French and Finch (2006), Chen (2007), Meade et al. (2008), Cheung and Lau (2012), Rutkowski and Svetina (2014). However, these procedures perform equally poorly when trying to detect violation of metric MI as given by PFL.

[7] This is confirmed by a corresponding Monte Carlo study, where the second group's data generating process was equal to that of the first group and where metric MI was (wrongly) rejected in 55 out of 1,000 cases.

For $\widetilde{\Lambda}_2 := 7/4\Lambda_2$ and $\widetilde{\Phi}_2 := (4/7)^2\Phi_2$, we obviously have $\widetilde{\Lambda}_2\widetilde{\Phi}_2\widetilde{\Lambda}_2' = \Lambda_2\Phi_2\Lambda_2'$. Therefore, due to formula (3), Λ_1, Λ_2 and Φ_1, Φ_2 on the one hand and $\Lambda_1, \widetilde{\Lambda}_2$ and $\Phi_1, \widetilde{\Phi}_2$ on the other hand lead to the same covariance matrix for the observed variables and thus to the same data generating process (dgp) for the observed variables. Therefore, even if we observed infinitely many data and were able to fully and correctly infer the observed variables' distribution, we could not infer whether the data were generated by the original setup or from the modified one. Thus, building on observed data only, it is impossible to decide whether data stem from the original quantities Λ_1, Λ_2 and Φ_1, Φ_2 or from the modified quantities given by $\Lambda_1, \widetilde{\Lambda}_2$ and $\Phi_1, \widetilde{\Phi}_2$. In other words, the original setup is empirically indistinguishable from the alternative one. However, while the original setup exhibits proportional but non-invariant factor loadings, the modified setup features invariant factor loadings! Taken together, we have two different setups which imply the same dgp for the observed data, one with metric MI and one with PFL, but without metric MI. This is exactly the reason behind the impossibility of telling apart PFL from metric MI, not only in the above example, but also more generally.

In order to see why PFL is in general empirically indistinguishable from metric MI, recall the defining condition for PFL from equation (4): the loadings in the different groups only differ by some multiplicative constants $c_1, \ldots, c_G$: $c_1\Lambda_1 = \ldots = c_G\Lambda_G$. We may thus modify the latent variables' loadings and covariances and construct $\widetilde{\Lambda}_1 := c_1\Lambda_1, \ldots, \widetilde{\Lambda}_G := c_G\Lambda_G$ and $\widetilde{\Phi}_1 := (1/c_1)^2\Phi_1, \ldots, \widetilde{\Phi}_G := (1/c_G)^2\Phi_G$, to arrive at a reformulation of the model with invariant loadings, but identical dgp, due to $\widetilde{\Lambda}_g\widetilde{\Phi}_g\widetilde{\Lambda}_g' = \Lambda_g\Phi_g\Lambda_g'$ for all groups $g = 1, \ldots, G$. Thus, every model satisfying PFL can be restated as a model satisfying metric MI, without changing the underlying data generating process. Put differently, for every PFL model there exists an empirically indistinguishable alternative model which fulfils metric MI.

The consequences of the empirical indistinguishability of PFL and metric MI are severe: every statistical test designed for testing metric MI at a significance level of, say, $\alpha = 5\%$, will be limited to detect PFL with a probability of at most α. This is necessarily so because the probability of (wrongly) rejecting metric MI is bounded by α under dgp's of metric MI. However, as every dgp under PFL is identical to one under metric MI, this implies that the probability of (correctly) rejecting metric MI under PFL is bounded by the same constant α. Put differently, all tests for metric MI suffer from no power to detect PFL. Or, rephrasing again, testing for metric MI is actually tantamount to testing for PFL.

This finding is a more precise formulation of several statements in the literature: for instance, Raykov et al. (2012) rightfully argue that the tests for metric MI are not 'complete', by which they mean that the tests do not check 'complete' invariance, but only whether the ratios of all other indicators' loadings to that of the referent indicator are invariant.[8] This

[8]Similar considerations can be found in Appendix B of Cheung and Rensvold (1999).

condition, i.e. invariance of these ratios, is actually easily seen to be equivalent to PFL.

Overall, therefore, one may summarize as follows: procedures designed for testing metric MI are actually testing PFL only, i.e. they will not detect non-invariant loadings according to PFL, but only violations of PFL. Therefore, they do not have power to detect PFL, but only power to detect violations of PFL, with power increasing the more the dgp departs from proportional factor loadings. This is the reason driving the above discussed intriguing and puzzling results found by Meade and Lautenschlager (2004), Meade and Bauer (2007), Chen (2007), Yoon and Millsap (2007), Chen (2008), French and Finch (2008), and Whittaker and Khojasteh (2013).

4 Testing for Metric Measurement Invariance

We will now discuss how testing for metric MI is compromised by the fact that one actually only tests for PFL. First of all, let us mention that, implicitly, this fact has already been hinted at by several authors, see e.g. Cheung and Rensvold (1999) and Raykov et al. (2012) and the references given therein. In particular, Cheung and Rensvold (1999) state rightfully: 'All such procedures embody a tacit *assumption* of invariance, even though the purpose of the procedures described above is to *test* for invariance.'[9]

In principle, there are two distinct cases: when a test for metric MI leads to rejection of the null hypothesis, then the data speaks against both metric MI and PFL. In such a situation, researchers have clear evidence that metric MI probably is violated by the data.
However, the situation is much more complicated when a test for metric MI fails to reject the null hypothesis: in this case, the data could possibly stem from a model with metric MI, but as well from a model with PFL only. How should researchers then decide between metric MI and PFL? Reconsidering the example discussed above: on what grounds should we decide whether the PFL specification with $\Lambda_1 = (.7, .9, .5, .6, .8, .3)'$, $\Lambda_2 = (0.4, 0.5143, 0.3429, 0.2857, 0.4571, 0.1714)'$ and factor variances of $\Phi_1 = 1$ and $\Phi_2 = 1.3$ is to be preferred over the empirically indistinguishable metric MI specification with $\Lambda_1 = (.7, .9, .5, .6, .8, .3)'$, $\Lambda_2 = (.7, .9, .5, .6, .8, .3)'$ and factor variances of $\Phi_1 = 1$ and $\Phi_2 = (4/7)^2 \cdot 1.3 = 0.4245$?[10] One might argue that the latter is to be preferred because it is somewhat more plausible that the loadings are actually invariant and not only proportional. However, one might also argue that the variance of the latent variable under consideration should be identical in both groups, in particular if there is evidence for this hypothesis, be it from theory or other empirical studies. In that case, one might want to

[9] With italics as in Cheung and Rensvold (1999).

[10] When considering alternative specification, the error terms' variances always stay unchanged.

consider a third empirically indistinguishable alternative, given by $\Lambda_1 = (.7, .9, .5, .6, .8, .3)'$, $\Lambda_2 = 4/7 \cdot \sqrt{1.3} \cdot (.7, .9, .5, .6, .8, .3)' = (0.4561, 0.5864, 0.3258, 0.3909, 0.5212, 0.1955)'$ and factor variances of $\Phi_1 = \Phi_2 = 1$. To emphasize again, this decision can *not* be taken on statistical grounds alone. All three model variants lead to identical distributions of the observed variables, and only reasons from outside of statistics may lead us to a decision in favor of a model with invariant loadings but non-invariant latent variances, or in favor of a model with non-invariant proportional loadings but invariant latent variances, or in favor of a model where neither loadings nor latent variances are invariant.[11] While in one application, going for the invariant loadings may be appropriate due to information from outside of statistics, it might in another application be the right thing to opt for a model with proportional loadings and non-invariant latent variances.

It is important to notice that such a choice may have severe consequences: wrongly deciding in favor of metric MI might lead to the conclusion that latent variances differ across groups, although the differences stem from loadings that are only proportional. On the other hand, assuming that the latent variances are invariant may lead to wrong conclusions when they are in fact non-invariant. We will discuss these issues more deeply in the next section.

Finally, we want to point out again that the above discussed problems occur independently of what scaling method is used to identify the unconstrained model of configural and the constrained model of metric invariance, respectively. As already stated above, the results of testing for metric invariance do not depend on the scaling method, because χ^2-values and other fit indexes are not affected by the particular choice of the scaling method.[12] Thus, as long as one only wants to test for metric MI, there is no 'referent indicator problem', as the test's result does not depend on which indicator is chosen.[13]

5 Testing the Latent Variables' Variances

In the following, we explore how testing for equality of latent variances may be affected when factor loadings are not invariant, but only proportional. To this end, we consider five different dgp's of which two are characterized by metric MI, while the other three have non-invariant factor loadings according to PFL. The corresponding parameters are described in Table 1:[14]

[11] The alignment method of Asparouhov and Muthén (2014) may be interpreted as postulating that in such a case, one should opt for the model satisfying metric MI.

[12] This independence of the scaling method has also been confirmed by Johnson et al. (2009) by a corresponding Monte Carlo study.

[13] However, when trying to detect non-invariant indicators, the choice of the referent indicator may matter indeed, see below.

[14] For all dgp's given in Table 1, error term variances are invariant over groups, to save space, they are presented in Table 1 only once.

	MI_EV	MI	Prop	Prop_SEV	Prop_EV
λ_{11}	0.7000	0.7000	0.7000	0.7000	0.7000
λ_{21}	0.9000	0.9000	0.9000	0.9000	0.9000
λ_{31}	0.5000	0.5000	0.5000	0.5000	0.5000
λ_{41}	0.6000	0.6000	0.6000	0.6000	0.6000
λ_{51}	0.8000	0.8000	0.8000	0.8000	0.8000
λ_{61}	0.3000	0.3000	0.3000	0.3000	0.3000
Φ_1	1.0000	1.0000	1.0000	1.0000	1.0000
λ_{12}	0.7000	0.7000	0.4000	0.4000	0.4000
λ_{22}	0.9000	0.9000	0.5143	0.5143	0.5143
λ_{32}	0.5000	0.5000	0.2857	0.2857	0.2857
λ_{42}	0.6000	0.6000	0.3429	0.3429	0.3429
λ_{52}	0.8000	0.8000	0.4571	0.4571	0.4571
λ_{62}	0.3000	0.3000	0.1714	0.1714	0.1714
Φ_2	1.0000	1.3000	1.3000	3.0000	1.0000
Θ_1	0.7000	0.7000	0.7000	0.7000	0.7000
Θ_2	1.2000	1.2000	1.2000	1.2000	1.2000
Θ_3	0.4000	0.4000	0.4000	0.4000	0.4000
Θ_4	0.5000	0.5000	0.5000	0.5000	0.5000
Θ_5	0.9000	0.9000	0.9000	0.9000	0.9000
Θ_6	0.2000	0.2000	0.2000	0.2000	0.2000

Table 1: Parameters for different models. Error term variances invariant across groups.

the data generating process called 'Prop' is exactly the one that has already been discussed above, it serves as a basic dgp from which the other dgp's are derived. For instance, the dgp 'MI' is very similar to 'Prop', with the only difference being that for 'MI', the factor loadings in the second group are identical to those in the first group. 'MI_EV' builds on 'MI', the only difference to 'MI' is that 'MI_EV' has equal latent variances in both groups. On the other hand, 'Prop_SEV' and 'Prop_EV' differ from 'Prop' only with respect to the latent variance, which is even more invariant across groups for 'Prop_SEV' and perfectly invariant for 'Prop_EV'.

For all these dgp's, we simulated 1,000 replications with a sample size of 500 in each group. In line with the above, testing for metric MI via the likelihood ratio χ^2-difference test at a significance level of 5% lead to rejecting metric MI only in 57, 55, 54, 54, and 58 cases for 'Prop', 'MI', 'MI_EV', 'Prop_SEV', and 'Prop_EV', respectively. In the vast majority of cases, one would thus have concluded that the data fulfil metric MI, these conclusions being correct only for the dgp's 'MI' and 'MI_EV', but unknowingly erroneous for the three cases of PFL, where the data stem from 'Prop', 'Prop_SEV', and 'Prop_EV'. In all these cases where metric MI is not rejected, researchers typically might want to test whether the latent

variable's variance is invariant over the two groups, as one is often interested in detecting whether the underlying factor's variation does differ between groups or not.

The results of the corresponding tests, using a significance level of 5%, are as follows. For 'MI', equality of latent variances is (rightfully) rejected for 621 out of the 945 cases: thus, for 'MI', the test can detect with a power of approximately 65.7% that the latent variables' variances differ across groups. For 'MI_EV', equality of latent variances is rejected in only 40 out of 946 cases, showing that such a wrong conclusion happens only with a probability of roughly 4.23%. For 'Prop', equality of latent variances was (rightfully) rejected in 943 out of 943 cases, resulting in perfect power to detect the non-invariant latent variances. For 'Prop_SEV', however, where the discrepancy between the latent variances in the two groups is much more pronounced, equality of latent variances was rejected only in 44 out of 946 cases, i.e. with a probability of only approximately 4.65%. Put differently, although latent variances under 'Pro_SEV' are very distinct, this difference is detected with a probability smaller than the significance level of the test of 5%. In other words, there is no power at all to detect the diverging latent variances under 'Prop_SEV', the probability of wrongly accepting equal the hypothesis of latent variances being larger than 95%.[15] For 'Prop_EV', on the other hand, where latent variances coincide in the two groups, the hypothesis of equal variances is wrongly rejected in 942 out of 942 cases.[16]

Summing up, we can state that testing for equality of latent variances may be severely distorted when the data do not stem from metric MI, but only from PFL. First of all, under PFL, the non-invariance of the factor loadings will typically remain undiscovered. As a consequence, the probability of wrongly accepting the hypothesis of no variation of latent variances across groups may be very large, even if the variances differ drastically across groups. Furthermore, it is also possible that the test will be fooled to wrongly reject the hypothesis of equal variances with a high probability, although latent variances are actually invariant. In a nutshell: testing for equality of latent variances is severely compromised when factor loadings are proportional, but not invariant across groups.

[15]The reason underlying the failure to detect the diverging latent variances is that under 'Prop_SEV', the dgp in the second group is empirically indistinguishable from an alternative one where the loadings are $7/4 \cdot (0.4, 0.5143, 0.2857, 0.3429, 0.4571, 0.1714) = (0.7, 0.9, 0.5, 0.6, 0.8, 0.3)$, i.e. equal to the loadings in the first group, and the factor variance is $3 \cdot (4/7)^2 \approx 0.9796$.

[16]The reason for the large probability of wrongly rejecting the hypothesis of equal factor variances is that under 'Prop_EV', the dgp in the second group is empirically indistinguishable from an alternative one where the loadings are $7/4 \cdot (0.4, 0.5143, 0.2857, 0.3429, 0.4571, 0.1714) = (0.7, 0.9, 0.5, 0.6, 0.8, 0.3)$, i.e. equal to the loadings in the first group, and the factor variance is $1 \cdot (4/7)^2 \approx 0.3265$.

	Ind. 1	Ind. 2	Ind. 3	Ind. 4	Ind. 5	Ind. 6
Loading in Group 1	0.30	0.10	0.20	0.40	0.60	0.20
Loading in Group 2	0.60	0.20	0.40	0.80	0.60	0.20
Loading in Group 3	0.60	0.10	0.20	0.40	0.30	0.10
Error Variance	0.20	0.30	0.10	0.15	0.25	0.05

Table 2: Basic Example on partial MI. Error term variances invariant across groups, latent variances $\Phi_1 = \Phi_2 = \Phi_3$ equal to 4 in all groups.

6 Analyzing Partial Metric Measurement Invariance

In empirical applications, researchers often find that their data are such that the hypothesis of metric measurement invariance is untenable. Although this might indicate that latent factors measure different things in different groups, one often proceeds by analyzing partial measurement invariance, by which models are meant where the loadings of only some, but not necessarily all indicators are supposed to be invariant. In the literature, a lot of attention has been paid to this topic, starting almost thirty years ago with Byrne et al. (1989), but also very recently, see e.g. Raykov et al. (2013), Yoon and Kim (2014), Jung and Yoon (2016), and Jung and Yoon (2017). In the meantime, the corresponding problem of 'Locating the Violation of Invariance' has been named as one of four unsolved problems in studies of factorial invariance by Millsap (2005).

In order to study the problems associated with partial invariance, we will have a detailed look at the following example, where in three groups, a single factor is measured by six indicators, see Table 2: none of the indicators has completely invariant factor loadings, as for every indicator the loading in one group differs from the loading in the other two groups. The latent variable's variance was set to 4 in all groups and error term variances were also invariant across groups, with values as given in Table 2.

Notice that the distribution of the observed variables does not change when we rescale the factor in the first group according to Table 3: in this case, the loadings in the first group become twice as large as in the orginal setup, and the latent variance in the first group changes from 4 to unity. In this empirically indistinguishable case, there is exactly one indicator with invariant factor loadings, namely Indicator 1. Analogously, one may alter the basic setup by changing the latent factor's scaling in the second group according to Table 4: in this case, the loadings in the second group are only half as large as in the original setup, and the latent variance correspondingly changes to 16. This second modification is again empirically indistinguishable from the original setup, and strangely enough, now indicators 2, 3, and 4 display invariant loadings across groups. Finally, Table 5 shows yet another modification resulting in an empirically indistinguishable model: it is constructed by doubling

	Ind. 1	Ind. 2	Ind. 3	Ind. 4	Ind. 5	Ind. 6
Loading in Group 1	0.60	0.20	0.40	0.80	1.20	0.40
Loading in Group 2	0.60	0.20	0.40	0.80	0.60	0.20
Loading in Group 3	0.60	0.10	0.20	0.40	0.30	0.10
Error Variance	0.20	0.30	0.10	0.15	0.25	0.05

Table 3: Modification A on partial MI. Error term variances invariant across groups, latent variances equal to $\Phi_1 = 1$, $\Phi_2 = 4$, $\Phi_3 = 4$, respectively.

	Ind. 1	Ind. 2	Ind. 3	Ind. 4	Ind. 5	Ind. 6
Loading in Group 1	0.30	0.10	0.20	0.40	0.60	0.20
Loading in Group 2	0.30	0.10	0.20	0.40	0.30	0.10
Loading in Group 3	0.60	0.10	0.20	0.40	0.30	0.10
Error Variance	0.20	0.30	0.10	0.15	0.25	0.05

Table 4: Modification B on partial MI. Error term variances invariant across groups, latent variances equal to $\Phi_1 = 4$, $\Phi_2 = 16$, $\Phi_3 = 4$, respectively.

the original loadings in the third group, and accordingly reducing the latent variance from 4 to 1. Again, this modification is empirically indistinguishable from the original setup, but now the indicators 5 und 6 are characterized by invariant factor loadings.

From the above, it becomes clear that, based on data only, it is impossible to infer which indicators are truly invariant: the above discussed models constitute four cases that cannot be told apart empirically, but sometimes there are no invariant loadings at all, sometimes only the first indicator's loadings are invariant, sometimes those of indicators 2, 3, and 4, and sometimes those of indicators 5 and 6.

Although it is impossible to infer the invariant factor loadings from data only, we still can learn a lot from properly studying given data. In order to exemplify this, we simulated 1,000 replications of the above described example and estimated the corresponding configural model using different scaling methods. In particular, we used all indicators in turn as referent indicators by enforcing the corresponding loading to be equal to unity in all groups. Furthermore, we additionally estimated the model using the effects coding method. Table 6 shows

	Ind. 1	Ind. 2	Ind. 3	Ind. 4	Ind. 5	Ind. 6
Loading in Group 1	0.30	0.10	0.20	0.40	0.60	0.20
Loading in Group 2	0.60	0.20	0.40	0.80	0.60	0.20
Loading in Group 3	1.20	0.20	0.40	0.80	0.60	0.20
Error Variance	0.20	0.30	0.10	0.15	0.25	0.05

Table 5: Modification C on partial MI. Error term variances invariant across groups, latent variances equal to $\Phi_1 = 4$, $\Phi_2 = 4$, $\Phi_3 = 1$, respectively.

	I1	I2	I3	I4	I5	I6	EC
λ_{11}	1.0000	3.0406	1.5007	0.7509	0.5001	1.5018	0.9998
λ_{21}	0.3351	1.0000	0.5021	0.2513	0.1674	0.5026	0.3343
λ_{31}	0.6680	2.0282	1.0000	0.5011	0.3337	1.0021	0.6672
λ_{41}	1.3340	4.0501	1.9996	1.0000	0.6663	2.0009	1.3322
λ_{51}	2.0029	6.0816	3.0021	1.5022	1.0000	3.0043	2.0002
λ_{61}	0.6671	2.0260	1.0000	0.5003	0.3332	1.0000	0.6663
ϕ_1	0.3598	0.0409	0.1602	0.6382	1.4388	0.1596	0.3596
λ_{12}	1.0000	3.0142	1.4993	0.7505	1.0000	3.0023	1.2857
λ_{22}	0.3333	1.0000	0.4995	0.2500	0.3331	1.0003	0.4282
λ_{32}	0.6674	2.0110	1.0000	0.5007	0.6672	2.0032	0.8579
λ_{42}	1.3330	4.0168	1.9980	1.0000	1.3327	4.0010	1.7135
λ_{52}	1.0006	3.0151	1.4998	0.7507	1.0000	3.0033	1.2862
λ_{62}	0.3334	1.0047	0.4997	0.2501	0.3333	1.0000	0.4285
ϕ_2	1.4385	0.1603	0.6405	2.5542	1.4395	0.1599	0.8700
λ_{13}	1.0000	6.1345	3.0005	1.5000	2.0032	6.0054	2.1178
λ_{23}	0.1657	1.0000	0.4970	0.2484	0.3317	0.9948	0.3505
λ_{33}	0.3338	2.0470	1.0000	0.5006	0.6685	2.0043	0.7068
λ_{43}	0.6673	4.0910	2.0015	1.0000	1.3361	4.0058	1.4127
λ_{53}	0.5001	3.0660	1.5002	0.7499	1.0000	3.0024	1.0587
λ_{63}	0.1670	1.0243	0.5010	0.2504	0.3345	1.0000	0.3536
ϕ_3	1.4394	0.0401	0.1605	0.6406	0.3603	0.0402	0.3209

Table 6: Results of Monte Carlo study on partial MI. 'I1' through 'I6' refers to fixed marker scaling using indicators 1 through 6 as referent indicators, while 'EC' denotes 'effects coding'. The table reports the averaged estimates of the corresponding model parameters, the average being taken over 1,000 runs of Model 2.

the corresponding results.

Using the first indicator as referent indicator, it seems that only this first indicator's loadings are invariant across groups. However, as the corresponding loadings were fixed at unity from the outset, this is no conclusive evidence of the first indicator's invariance, as has been pointed out namy times in the literature. The more interesting fact here is that no other indicator appears to have invariant factor loadings when the first indicator is used as referent indicator. In this sense, the corresponding results can be interpreted as reflecting modification A given in Table 3.

When indicators 2, 3, or 4 are used as referent indicators, Table 6 conveys the impression that only the loadings of these indicators are invariant. Again, this is completely in line with the corresponding modification B given in Table 4. In the literature, the set $\{2, 3, 4\}$ has sometimes been called an 'invariant set' of indicators, see e.g. Rensvold and Cheung (1998). In light of the previous discussion, we argue that a better name would be 'proportional set',

as the loadings of indicators 2, 3, and 4 behave proportionally over the groups, but from this, one must not conclude that they are actually invariant.

Upon using indicators 5 or 6 as referent indicators, Table 6 suggests that only indicators 5 and 6 are characterized by invariant loadings, completely in line with modification C given in Table 5. Similar to above, we propose to call $\{5, 6\}$ a proportional set of indicators.

Finally, when effects coding is used, Table 6 creates the impression that no indicator possesses invariant loadings, a conclusion which is in line with the original setup as given by Table 2.

Overall, the simulation shows that it is well possible to detect the 'proportional sets', i.e. those subsets of indicators whose loadings behave proportionally across groups. One method to find these proportional sets is to estimate the model using in turn all indicators as referent indicators and then check which indicators appear to have invariant loadings. Another possibility would be to estimate the model only once, using an arbitrary scaling method, and then compute the ratios of one indicator's loadings over another indicator's loadings. For instance, the ratios of the second over the third indicator's loadings is approximately 0.5 in all groups in Table 6, regardless of which scaling method is used, in line with the corresponding ratios in Tables 2-5.

However, even though one may identify the proportional sets, it still is impossible to learn from data only which, if any, of these sets actually corresponds to invariant factor loadings. In applications, such a conclusion cannot be based purely on the data, but must be drawn from other reasons like theory or previous empirical results.

7 Conclusion

In this paper, we have shed new light on the topic of metric measurement invariance. In particular, we have introduced the notion of proportional factor loadings (PFL), which is given when loadings across groups coincide only up to multiplicative constants. We found that it is impossible to tell apart metric MI from PFL and that testing for metric MI is actually tantamount to testing for PFL. Researchers therefore should be aware of the fact that when metric MI is not rejected, an additional assumption is necessary to proceed from PFL to metric MI. This additional assumption must be backed by theory or earlier empirical results, it can not be justified on the data at hand.

Researchers should also be aware of the fact that tests for invariance of latent variances may be severely distorted and lead to completely wrong conclusions when the data generating process only satisfies PFL, but not metric MI.

Finally, proportional factor loadings are also important to understand some phenomena

related to partial MI. With respect to partial MI, the most important insight of this paper is that from data alone, it is not possible to infer which indicators's loadings are invariant and which are not. Instead, empirical data only allows to find subsets of indicators whose loadings are proportional across groups: in analogy to the term 'invariant sets' used in the literature, these sets are called 'proportional sets'. Thus, while statistics, by detecting proportional sets, can help the researcher to find out which indicators behave similar by loading proportionally across groups, it is up to the researcher to decide, guided by theory or earlier empirical research, which of these proportional sets may be assumed to be an 'invariant set'.

References

Asparouhov, T. and Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(1):1–14.

Byrne, B. M., Shavelson, R. J., and Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *The Psychological Bulletin*, 105:456–466.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3):464–504.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? the impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, 95(5):1005–1018.

Cheung, G. W. and Lau, R. S. (2012). A direct comparison approach for testing measurement invariance. *Organizational Research Methods*, 15(2):167–198.

Cheung, G. W. and Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, 25(1):1–27.

Cheung, G. W. and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A multidisciplinary Journal*, 9(2):233–255.

French, B. F. and Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(3):378–402.

French, B. F. and Finch, W. H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling: A Multidisciplinary Journal*, 15(1):96–113.

Johnson, E. C., Meade, A. W., and DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(4):642–657.

Jung, E. and Yoon, M. (2016). Comparisons of three empirical methods for partial factorial invariance: Forward, backward, and factor-ratio tests. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(4):567–584.

Jung, E. and Yoon, M. (2017). Two-step approach to partial factorial invariance: Selecting a reference variable and identifying the source of noninvariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(1):65–79.

Little, T. D., Slegers, D. W., and Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in sem and macs models. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(1):59–72.

Meade, A. W. and Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(4):611–635.

Meade, A. W., Johnson, E. C., and Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, 93(3):568–592.

Meade, A. W. and Lautenschlager, G. J. (2004). A monte-carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 11(1):60–72.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4):525–543.

Millsap, R. E. (2005). Four unresolved problems in studies of factorial invariance. In Maydeu-Olivares, A. and McArdle, J. J., editors, *Contemporary psychometrics: a festschrift for Roderick P. McDonald*, Multivariate applications book series. Erlbaum, Mahwah, NJ.

Putnick, D. L. and Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41:71 – 90.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raykov, T., Marcoulides, G. A., and Li, C.-H. (2012). Measurement invariance for latent constructs in multiple populations. *Educational and Psychological Measurement*, 72(6):954–974.

Raykov, T., Marcoulides, G. A., and Millsap, R. E. (2013). Factorial invariance in multiple populations. *Educational and Psychological Measurement*, 73(4):713–727.

Rensvold, R. B. and Cheung, G. W. (1998). Testing measurement models for factorial invariance: A systematic approach. *Educational and Psychological Measurement*, 58(6):1017–1034.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.

Rutkowski, L. and Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, 74(1):31–57.

Whittaker, T. A. and Khojasteh, J. (2013). A comparison of methods to detect invariant reference indicators in structural equation modelling. *International Journal of Quantitative Research in Education*, 1(4):426–443.

Yoon, M. and Kim, E. S. (2014). A comparison of sequential and nonsequential specification searches in testing factorial invariance. *Behavior Research Methods*, 46(4):1199–1206.

Yoon, M. and Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A monte carlo study. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3):435–463.